DOCUMENT RESUME

ED 126 816 HE 008 131

AUTHOR Carroll, Stephen J.; Relles, Daniel A.
TITLE A Bayesian Model of Choice Among Higher Education Institutions.
INSTITUTION Rand Corp., Santa Monica, Calif.
SPONS AGENCY Lilly Endowment, Inc., Indianapolis, Ind.; National Inst. of Education (DHEW), Washington, D.C.
REPORT NO R-2005-NIE/LE
PUB DATE Jun 76
GRANT NIE-G-74-0038
NOTE 41p.
AVAILABLE FROM Rand, Santa Monica, California 90406

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Bayesian Statistics; Comparative Analysis; Data Analysis; *Data Bases; *Higher Education; Models; *Prediction; *Probability Theory; *Student Characteristics
IDENTIFIERS *Conditional Logit Analysis

ABSTRACT

Examined are methodologies for modeling students' choices among higher education institutions. A statistical technique called "conditional logit analysis" is applicable to the problem studied. These applications are reviewed and certain weaknesses inherent in the approach are pointed out. Alternative approaches are offered, based on the use of Bayes' Theorum, which is easier to use, more flexible, and less expensive to apply. In empirical tests, it is also observed to have greater predictive power than conditional logit analysis. (Author)

**********
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
**********

ED126816

# A BAYESIAN MODEL OF CHOICE AMONG HIGHER EDUCATION INSTITUTIONS

PREPARED UNDER GRANTS FROM THE NATIONAL INSTITUTE OF EDUCATION AND LILLY ENDOWMENT, INCORPORATED

STEPHEN J. CARROLL
DANIEL A. RELLES

JUNE 1976
R-2005-NIE/LE

HE 20 8131

2

Rand
A NONPROFIT CORPORATION

The research reported herein was performed in part pursuant to Grant No. NIE-G-74-0038 from the National Institute of Education, Department of Health, Education, and Welfare and in part to a grant from Lilly Endowment, Incorporated. The opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education or of Lilly Endowment, Incorporated, and no official endorsement by the National Institute of Education or by Lilly Endowment, Incorporated should be inferred.

Published by The Rand Corporation.

# A BAYESIAN MODEL OF CHOICE AMONG HIGHER EDUCATION INSTITUTIONS

PREPARED UNDER GRANTS FROM THE NATIONAL INSTITUTE OF EDUCATION AND LILLY ENDOWMENT, INCORPORATED

STEPHEN J. CARROLL
DANIEL A. RELLES

JUNE 1976
R-2005-NIE/LE

Rand
SANTA MONICA, CA. 90406

## PREFACE

This report was prepared with support from the National Institute of Education and the Lilly Endowment, Inc. The purpose of the research was to examine methodologies for modeling students' choices among higher education institutions.

A statistical technique called "conditional logit analysis" has recently been popularized; its applications include exactly the problem studied here. The authors review these applications and point out certain weaknesses inherent in the approach. They then offer an alternative approach, based on the use of Bayes' Theorem, which is easier to use, more flexible, and less expensive to apply. In empirical tests, it was also observed to have greater predictive power than conditional logit analysis.

The authors are grateful to Rand colleagues Bryan C. Ellickson, Gus W. Haggstrom, and John J. McCall for valuable comments on an earlier draft of this report.

## SUMMARY

This study revisits a problem that has received considerable attention in recent years: modeling students' choices among institutions of higher education. We offer a methodological approach to the problem which obviates some of the technical and methodological difficulties encountered in previous studies, where the primary tool of analysis has been "conditional logit." We demonstrate our approach with data from the SCOPE 1966 survey of high school seniors and compare our results to those obtained in other analyses of the SCOPE data.

We regard the SCOPE data as drawn from a population described by a joint density $P(i,j)$, where $i$ identifies a particular student and $j$ a particular institution. The problem is to obtain a parametric model for $P(j|i)$, the probability that student $i$ chooses institution $j$. The conditional logit approach uses a maximum likelihood technique to estimate $P(j|i)$ directly, whereas we suggest a two-stage procedure in which the parameters of $P(i|j)$ are estimated via ordinary linear regression, then Bayes' Theorem is used to obtain $P(j|i)$. The regression models describe student ability, income, and distance from home as functions of the characteristics of chosen institutions. In using Bayes' Theorem, we assume that the prior probability of choosing a given institution depends on its size.

We apply our model to the problem of predicting the distribution of students among certain homogeneous categories of institutions. We find that the deviations between predicted and actual distributions are quite small and that the predictive power of our model is substantially greater than that of alternative models which used the conditional logit methodology to analyze the same data set.

Conditional logit studies of individual choice behavior in a variety of areas have recently appeared in the literature. Our results suggest that the Bayesian formulation is a viable alternative. Questions of predictive power aside, the Bayesian methodology is easier to use, offers much greater flexibility, and is less expensive to apply. Thus, even in cases where theoretical considerations might suggest the alternative approach, the Bayesian methodology would be a useful adjunct in the exploratory stages of research.

CONTENTS


| | | |
|---|---|---|
| PREFACE | | iii |
| SUMMARY | | v |
| Section | | |
| I. | INTRODUCTION | 1 |
| II. | THE CONDITIONAL LOGIT APPROACH | 3 |
| | The Formal Structure | 3 |
| | Previous Studies | 5 |
| | Limitations of the Approach | 9 |
| III. | A BAYESIAN ALTERNATIVE TO CONDITIONAL LOGIT | 13 |
| | The Formal Structure | 13 |
| | Distribution of Student Characteristics | 14 |
| | Student Choice Probabilities | 15 |
| | Features of the Approach | 16 |
| IV. | MODEL ESTIMATION | 18 |
| | Constructing the Data Base | 18 |
| | Estimating the Distribution of Student Characteristics | 19 |
| V. | PREDICTIVE POWER | 25 |
| | RM's Simulations | 25 |
| | Bayesian Simulations | 25 |
| | Comparison of Results | 28 |
| VI. | CONCLUDING REMARKS | 33 |
| REFERENCES | | 35 |

## I. INTRODUCTION

This study revisits a problem that has received considerable attention in recent years: modeling students' choices among institutions of higher education. Our primary objective is to offer a methodological approach to the problem which obviates some of the technical and methodological difficulties encountered in previous studies. We demonstrate the approach with data from the SCOPE 1966 survey of high school seniors,[1] and compare our results to those obtained in other analyses of the SCOPE data.

Our point of departure is the recent work of Kohn, Manski, and Mundel [1] and Radner and Miller [2,3].[2] Both used a statistical estimation technique called "conditional logit" to analyze students' choices, given their characteristics.[3] The conditional logit approach overcomes many of the limitations of the other available approaches.[4] But it has important limitations of its own.

The technique has very demanding data requirements. The analyst must know the entire set of alternatives each student considered in making his choice. Second, the computational problems involved in maximizing the logit likelihood function are so severe as to limit both the flexibility one has in choosing the functional form of the relationships and the amount of exploratory analysis one can do. It is barely

---

[1] School to College: Opportunities for Postsecondary Education. This survey, conducted by The Center for Research and Development, University of California, Berkeley, is described in Sec. II.

[2] Radner and Miller [2] present the analysis. Many of the technical details, however, are reserved to a separately published technical supplement--Miller and Radner [3]. For simplicity in discussion, we will consistently refer to their joint work as Radner and Miller, using bracketed reference numbers to distinguish between the two.

[3] The conditional logit approach has been recently popularized by McFadden [4,5]. It is now being applied in a broad range of studies of individual decisions including choices among transportation modes [6] and occupations [7,8].

[4] Radner and Miller [2] provide a detailed critique of the approaches used in earlier studies and outline the advantages of the conditional logit technique.

feasible to write down a single model specified by theory and then to estimate parameters. It is not feasible to admit that the theory is weak, and thus that alternative formulations of independent variables, goodness of fit tests, analyses of residuals, etc., should be tried.

We view these difficulties as motivation for our own approach, which begins with two basic observations. First, if one is to predict a student's choice, given his characteristics, it seems reasonable that one should be able to say something about his characteristics, given his choice. Second there exists a readily applicable method to translate statements about characteristics, given choice, to statements about choice, given characteristics--Bayes' Theorem.

Thus, we regard the SCOPE data as drawn from a population described by a joint density $P(i,j)$, where $i$ identifies a particular student and $j$ a particular institution. The problem is to obtain a parametric model for $P(j|i)$, the probability that student $i$ chooses institution $j$. The conditional logit approach uses a maximum likelihood technique to estimate $P(j|i)$ directly, whereas we suggest a two-stage procedure in which the parameters of $P(i|j)$ are estimated via ordinary linear regression, then Bayes' Theorem is used to obtain $P(j|i)$. The regression models describe student ability, income, and distance from home as functions of the characteristics of chosen institutions. In using Bayes' Theorem, we assume that the prior probability of choosing a given institution depends upon its size.

Section II reviews the conditional logit approach, describes the data available from the SCOPE 1966 survey, and reviews the Kohn, Manski, and Mundel and the Radner and Miller studies, focusing on the problems they encountered in using the conditional logit approach. Our approach is described in Sec. III. In Sec. IV, we describe our empirical results in deriving the parameters of $P(i|j)$. Section V provides an investigation of the predictive power of our approach as compared to that of Radner and Miller. Some concluding remarks are presented in Sec. VI.

## II. THE CONDITIONAL LOGIT APPROACH

In this section, after briefly reviewing the formal structure of the conditional logit approach, we summarize the Radner and Miller and the Kohn, Manski, and Mundel studies, describing their data bases, indicating the variables they used, and giving their procedures for imputing students' "choice sets." The section concludes with a discussion of some of the problems they encountered.

### THE FORMAL STRUCTURE[5]

The conditional logit approach is predicated on the assumptions that the alternative an individual chooses is preferred to all other alternatives available to him and that his preferences can be expressed in the form of a function defined over the attributes of alternatives. Formally, let $C_i$ be the set of mutually exclusive alternatives available to the *i*th student; let $X_i$ be his characteristics; let $Z_{ij}$ be the *j*th alternative's attributes with respect to him; and let $U_i(Z_{ij})$ be a scalar-valued measure of his preference for the *j*th alternative. He is assumed to choose the *j*th alternative if and only if $U_i(Z_{ij}) \geq U_i(Z_{ik})$ for all k in $C_i$. If differences among individuals' preferences for a given set of attributes have a random component $\varepsilon_{ij}$, the *i*th individual's preference for the *j*th alternative can be written $U(X_i, Z_{ij}, \varepsilon_{ij})$.

For reasons of tractability, it is necessary to assume that U is linear in parameters with an additive disturbance:

$$U(X_i, Z_{ij}, \varepsilon_{ij}) = V(X_i, Z_{ij}) \cdot \theta + \varepsilon_{ij} , \qquad (1)$$

where V is a vector valued function, $\theta$ is the vector of parameters to be estimated, and $\varepsilon_{ij}$ is a scalar random variable. The choice of alternative j imples:

$$V(X_i, Z_{ij}) \cdot \theta + \varepsilon_{ij} \geq V(X_i, Z_{ik}) \cdot \theta + \varepsilon_{ik} , \quad \text{for all } k \in C_i ,$$

---

[5]This subsection summarizes the discussion provided by Kohn, Manski, and Mundel [1] of the conditional logit analysis technique.

or equivalently,

$$(V(X_i, Z_{ij}) - V(X_i, Z_{ik})) \cdot \theta \geq \varepsilon_{ik} - \varepsilon_{ij} , \quad \text{for all } k \in C_i . \qquad (2)$$

In order to estimate the parameters of (2), it is necessary to specify the joint probability distribution of the $\varepsilon_{ij}$. A probability distribution that leads to a tractable likelihood function is the Weibull distribution:

$$\text{Prob } (\varepsilon \leq T) = e^{-\alpha e^{-\beta T}} , \quad \alpha > 0 , \beta > 0 .$$

If $\varepsilon_{ij}$ and $\varepsilon_{ik}$ are independent and identically distributed with this distribution, it can be shown that

Prob (j chosen from $C_i$)

$$= \text{Prob } (\varepsilon_{ik} - \varepsilon_{ij} \leq (V(X_i, Z_{ij}) - V(X_i, Z_{ik})) \cdot \theta , \quad \text{for all } k \in C_i)$$

$$= \frac{1}{1 + \Sigma_{k \in C_i, k \neq j} \exp\left(-\beta(V(X_i, Z_{ij}) - V(X_i, Z_{ik})) \cdot \theta\right)} . \qquad (3)$$

The likelihood of the observed choices made by a set of n individuals is

$$L(\beta, \theta) = \prod_{i=1}^{n} \text{Prob } (j_i \text{ chosen from } C_i) , \qquad (4)$$

where $j_i$ is the $i$*th* individual's choice.

Function optimization procedures can be used to determine the maximum likelihood estimates of the product $\beta\theta$. Knowledge of $\theta$ up to this multiple is sufficient for all applications.

## PREVIOUS STUDIES

### Data

Radner and Miller (RM) and Kohn, Manski, and Mundel (KMM) use the SCOPE 1966 survey of high school seniors.[6] The survey includes approximately 34,000 students in 305 public and private high schools in four states--California, Illinois, Massachusetts, and North Carolina. The baseline data obtained include personal and family characteristics, postsecondary aspirations and expectations, plans for postsecondary education, and sources of funds for college expenses. The Academic Ability Test (AAT), similar to the Scholastic Aptitude Test (SAT), was given to most of the students. Both KMM and RM convert AAT scores to the equivalent SAT scores.

In spring 1967, the SCOPE researchers attempted to "locate" the students who had gone on to college. The institutions each student had listed as his first or second college choice (in the baseline survey) and the junior college nearest his home were queried. Students were sent postcards requesting information on their current activities, and their high school counselors were asked if they knew where the students had gone. In all, a collegiate enrollment of 17,199 students was established. It was assumed that the 16,741 students not "located" at a college had not gone on to college.[7]

Responses to follow-up surveys were obtained from 10,581 college-going students, 8,683 parents of college-going students, and 3,014 parents of students who had not gone on to college.[8] The follow-up data included students' postsecondary activities and, if they had gone on to college, their expenses and sources of funds. Parents were asked to provide their 1966 family income.

---

[6] High school freshmen were also surveyed in 1966, and followed for four years, but neither RM nor KMM used that part of the data base.

[7] While many "nongoers" were positively identified (by their response to the follow-up postcard), it is likely that some college-going students are included among them. The data set does not distinguish between known nongoers and students never located.

[8] The numbers of students and parents to whom follow-up efforts were directed have not been published; response rates to the follow-up surveys are unknown.

Nonresponses and "don't know" responses to the family income question on the student baseline instrument were frequent. Moreover, RM [3] examine the cases where a student (on the 1966 baseline questionnaire) and his family (on the 1967 parent questionnaire) provided independent (1966) family income estimates and found substantial discrepancies between the two. Assuming that parent-reported income is more accurate, both RM and KMM developed income prediction equations by regressing parent reported family income on students' responses on parental education, job status, occupation, and income.

KMM obtained most of their data on institutional attributes from the 1966 Institutional Domain File compiled by the American Council on Education [9]. This file provides information on the tuition and fees, faculty, programs, student characteristics, financial aid, etc., of colleges and universities. To obtain a measure of the distance between a student's home and a college, KMM coded the latitude and longitude of SCOPE high schools and of colleges and universities and computed the straight-line distance, in miles, between each high school/college pair.

RM compiled data on institutions' attributes from research reports, institution catalogues, or direct correspondence. Instead of using a distance measure, RM inspected road maps and classified an institution as being within commuting distance of a student if it appeared possible to drive from the student's high school to the institution within 50 minutes.

## Models

RM's choice model focused on two variables: the ratio of cost to family income and the product of the student's ability (his SAT score) and the college's quality (the average SAT score of freshmen attending the institution).[9] They assumed that the "cost" of not going on to college was zero and that the "quality" of the "no-go" option was the

[9]RM defined the cost of attending an institution within commuting distance to be tuition plus $100 (books and supplies) plus $180 (transportation costs). If the institution was beyond commuting distance, they defined cost as tuition plus $100 (books and supplies) plus $180 (miscellaneous costs of living away from home) plus the approximate price of a round trip air fare plus $900 (room and board).

average SAT score of the California SCOPE students who had not gone on to college.

KMM modeled students' decisions as a two-stage process. In the first stage, each student evaluates the collegiate alternatives available to him and identifies the most preferred. This evaluation is assumed to depend on some 15 variables: tuition, tuition squared, distance, room and board fees, the average SAT score of the students attending the college, the squared difference between the student's SAT score and the average SAT score of the students attending the college, the college's revenues per student, the number of different areas in which the college has degree-granting programs, the percentage of students residing on campus, an indicator of single sex institutions, and a series of dummy variables indicating college type--private four-year college, private two-year college, public university, public four-year college, and public two-year college.[10] In the second stage, the student decides whether the most preferred college alternative is sufficiently attractive to induce him to enroll. This evaluation depends on father's education, mother's education, sex, and the highest preference "score" imputed to any college in the student's choice set.[11]

Imputing the Choice Set

In principle, each student had the option of enrolling at any college or university that would accept him. And, in 1966, there were over 2,300 institutions of higher education in the country, many of which were not selective. Even the academically weak SCOPE students could have gained admission to literally hundreds of institutions. Computational constraints, however, preclude analysis with choice sets

---

[10] KMM developed a separate "commuter choice" model to predict whether or not a student would commute to a college. If the prediction was to commute, distance was set equal to the number of miles between home and college; for these students, the room and board variable was set equal to zero. If the prediction was to reside, distance was set equal to zero and the college's dormitory fee was used for room and board.

[11] The college a student attended is included in his choice set; but if the preference score imputed to some other college exceeds the imputed preference score of his chosen college, the higher score is used as the measure.

of this magnitude. Thus, both RM and KMM had to devise procedures for imputing a choice set of manageable size for each student.

RM argue that the alternatives confronting any student can be clustered into ten basic groups. The first corresponds to the "no-go" option; the remaining nine correspond to institutions falling into various cost-by-quality categories. Table 1 summarizes the kinds of institutions they assign to each category.

Table 1

RADNER-MILLER CHOICE SETS: COST AND QUALITY CATEGORIES OF INSTITUTIONS

| Quality Category[a] | Low Cost Category (Less than $600) | Medium Cost Category ($600-$2250) | High Cost Category ($2250+) |
|---|---|---|---|
| Low (Less than 480) | Public 2-yr colleges within commuting distance | Trade schools and private 2-yr colleges within commuting distance | Private colleges and universities |
| Medium (480-540) | Public 4-yr colleges within commuting distance | Public 4-yr colleges beyond commuting distance and low-tuition private colleges within commuting distance | Private colleges and universities |
| High (540+) | Public universities within commuting distance | Public universities not within commuting distance | Private colleges and universities |

SOURCE: Radner and Miller [3], p. 43.

[a]Measure of quality = average SAT score of all students attending the institution.

For each student, RM identified all institutions that would have admitted him, had he applied.[12] They then calculated the average cost

[12]RM consulted high school counselors, college catalogues, admissions officers, and state officials to obtain estimates of the minimum SAT score required for entrance to the public institutions in each state

and quality of the "available" institutions in each category. If a student went on to college, the cost and quality attributes of the institution he attended were substituted for the average attributes of the institutions in its category. Each student's choice set thus comprised the "no-go" option, the institution he attended (if he went on to college), and eight (nine if he did not go on to college) "representative" institutions whose attributes were the mean values of the attributes of the institutions available to him in the corresponding cost/quality category.[13]

KMM constructed each college-going student's choice set by randomly selecting institutions located within 200 miles of the student's high school and applying an admissions model to determine whether or not it was available to the student.[14] Single sex colleges serving the opposite sex and colleges located more than 60 miles from the student which lacked residency facilities were rejected. The process was continued until ten "available" institutions were identified or until the set of institutions within 200 miles was exhausted. The institution actually attended was added to the ten, or fewer, colleges so identified to form the student's choice set.

## LIMITATIONS OF THE APPROACH

### Choice of Choice Set

The conditional logit approach requires that each student's choice

---

and estimated an "admissions model" for each of 400 private institutions. They assumed that an institution would admit a student whose SAT score exceeded the score estimated to yield a 50 percent admission probability.

[13] RM do not mention weights; they presumably used unweighted mean cost and quality measures to represent the institutions in a cost/quality group.

[14] Unlike RM, who constructed separate models for each institution, KMM estimated a single, albeit more detailed, admissions model for all institutions. In constructing students' choice sets, KMM estimated the probability that the student would be admitted to a (randomly selected) college. Rejecting schools for which admissions probability was less than .25, they generated a random number on the unit interval and included the institution in the choice set if the random number was less than the estimated admissions probability.

set be completely specified. This forced both RM and KMM to develop a number of peripheral data imputation models relating to choice sets and admission criteria at the individual student level. These procedures proved to be very costly. Both RM and KMM had intended to examine the entire SCOPE sample, but had to cut back substantially on the number of students. RM eventually concentrated their analysis on two subsamples, each including about 375 of the roughly 34,000 SCOPE students. And KMM could examine only the students in Illinois and North Carolina.

The data so laboriously constructed are of little independent interest. Estimates of students' choice sets, institutions' admissions patterns, and students' residency/commuter choices are of value only as input to the estimation of the conditional logit parameters. The accuracy of the imputed data is also open to question. The KMM procedure for imputing choice sets is based on the implausible assumption that every institution within 200 miles of a student's high school is equally likely to have been considered. And their approach to estimating a student's admissibility to an institution clearly leads to imputation errors--an institution is included in the student's choice set when he would not have been admitted there, and conversely.

RM avoid the problem of identifying the specific institutions a student considered by assuming that the student chooses among "representative" institutions whose attributes are the mean values of the attributes of institutions in various categories. They further stratify institutions by the attributes which enter the model (cost and quality), ensuring that the within-category variance of the variables is small and that each category's "representative" institution is similar to other institutions within its category. Since the mean attributes within a category are somewhat insensitive to the inclusion or exclusion of any particular institution, the accuracy of their admissions models is less critical. But this procedure is impractical if the variables in the model depend on more than two or three institutional attributes. As the number of institutional attributes included in the model is increased, one must expand the stratification scheme (vastly increasing computation costs) or enhance the risk of imputation errors

(differences between the attributes of the institutions a student considered and the mean attributes of the institutions in the various categories).

Computational Problems

The maximum likelihood procedures used to estimate the parameters of a conditional logit model are very expensive, limiting the extent to which alternative functional forms or specifications of variables can be explored within the research budget. One of KMM's college choice runs, for example, required 840 CPU seconds on an IBM 370/168 to estimate the parameters of a 10-variable specification for about 3,100 students having about 30,200 choices.[15] Another run to estimate a 20-variable specification of their go/no-go model for about 7,100 students required 1,040 CPU seconds.

This limitation is particularly apparent in RM's work. Beyond the variables which entered their model (institutional cost and quality, and student income and ability), they wished to explore the influence of some 21 additional student variables[16] on students' college-going rates and patterns. The natural approaches to the problem--estimating alternative specifications of the model which incorporated the additional variables and testing their significance, or stratifying the students by levels of the variables and fitting the model for each strata--were precluded by the prohibitive costs (and small cell sizes). Instead, RM used their basic model to predict the distributions of students, stratified by the variables to be explored, among postsecondary outcomes. These distributions were then compared to the students' actual distributions to discover whether "improved" predictions were obtained by taking account of differences among students in terms of the variables. The computational limitations of the maximum likelihood approach thus imposed an extremely cumbersome approach to the exploration of alternative specifications of the model.

---

[15] Each student's choice set included his chosen college and 10 (or fewer) imputed alternatives.

[16] Student's sex and various measures of student's attitudes, aspirations, and expectations. See [2, p. 51] for a list of variables.

Problems of Omitted Variables

The formulation of the conditional logit model in terms of individuals' preferences limits the analysis to variables that have a behavioral interpretation. Institutional size, for example, does not readily fit in unless one contends that the differences in sizes of institutions reflect differences in the perceived utilities of size to potential students. Neither RM nor KMM were willing to do that; both implicitly assume that institutions are large or small only because their other attributes are relatively attractive to many or few students. But size is important; it reflects a number of institutional attributes, some of which cannot easily be measured: academic reputation, capacity constraints, recruiting efforts, quality of football teams, climate, recreational facilities, proximity to population centers, etc. Thus, there is reason to believe that the KMM and RM lists of behavioral variables are incomplete, and that the fitting process has compensated by putting larger (smaller) coefficients on those variables positively (negatively) correlated with size.

## III. A BAYESIAN ALTERNATIVE TO CONDITIONAL LOGIT

Bayes' Theorem provides an alternative approach to the problems of modeling individuals' choices which, we contend, alleviates many of the problems discussed above. This section develops a general theory for estimating the probability $P(j|i)$ that individual i chooses institution j. We then summarize our empirical approach to estimating the distribution of student characteristics, deferring detailed discussion to Sec. IV. We show how student choice probabilities can be derived from these empirical results, and conclude with a discussion of the advantages of the approach.

### THE FORMAL STRUCTURE

As above, let $X_i$ denote the *ith* individual's vector of characteristics, $Z_{ij}$ the *jth* institution's vector of attributes with respect to the *ith* individual. Our goal is to obtain a convenient parameterization for $P(i|j)$ in terms of $X_i$ and $Z_{ij}$.

We model $X_i$ as a transformed multivariate normal vector with mean $\mu_{ij} = \mu(Z_{ij})$ and covariance matrix $\Sigma_{ij} = \Sigma(Z_{ij})$. Thus, our basic assumption is that

$$\{T(X_i)|Z_{ij}\} \sim N(\mu_{ij}, \Sigma_{ij}) ,$$

where T is a real-valued vector function. Letting

$$Y_i = T(X_i) ,$$

we note then that $P(i|j)$ can be replaced in Bayes' formula by the function

$$f(Y_i|Z_{ij}) \propto |\Sigma_{ij}|^{-1/2} \exp\{-1/2(Y_i - \mu_{ij})' \Sigma_{ij}^{-1} (Y_i - \mu_{ij})\} . \quad (5)$$

A slightly more general class of models is obtained by assuming that $Y_i$ may be broken into subcomponents $Y_{1i}$ and $Y_{2i}$. $Y_{1i}$ is assumed to be multivariate normal, given $Y_{2i}$ and $Z_{ij}$, with a mean vector and covariance matrix that depends in an unspecified manner on $Y_{2i}$ and $Z_{ij}$; $Y_{2i}$ is assumed to be multivariate normal, its parameters dependent only on $Z_{ij}$. The function f might then be factorized as follows:

$$f(Y_i|Z_{ij}) = f_1(Y_{1i}|Y_{2i}, Z_{ij})\ f_2(Y_{2i}|Z_{ij}) , \qquad (6)$$

where $f_1$ and $f_2$ are multivariate normal densities, as in Eq. (5).

DISTRIBUTION OF STUDENT CHARACTERISTICS

In our empirical work, we investigated probability distributions whose densities f could be factored as in Eq. (6). We tried transformations T that were simple, conditioned on location of high school ($Y_2$) in modeling other student characteristics ($Y_1$), and assumed that means $\mu_{ij}$ were linear in institutional attributes and that covariance matrices $\Sigma_{ij}$ were constant within groups of institutions. Thus, we were able to estimate parameters of the distributions of characteristics using ordinary linear regression.

We confined our attention to students who went on to college. Although the theory could just as easily have handled the nongoers as an additional category, we felt that it would lighten our load considerably to omit them and that it would still be possible to make direct comparisons with other studies.

Since our objective was to obtain the probability distribution of characteristics, it seemed practical (and prudent) to choose only a few important ones. KMM and RM stressed the importance of such student characteristics as ability, family income, and location of high school. Similarly, they focused on a small subset of institutional attributes: type of institution (public or private, two- or four-year), cost and location. These variables were available in the SCOPE and Institutional Domain File data bases.

In estimating the parameters of the distribution of student characteristics, we concluded with a simple model in which students were

stratified by state of residence and sex: eight categories in all. Within strata, student ability (measured by the sum of verbal and mathematical AAT test scores) was regressed on institutional quality (measured by the mean SAT test scores of students attending the institution); the logarithm of family income was regressed on the estimated cost of attending the institution; and the logarithm of the distance between the student's home and institution was regressed on a constant. We examined the residuals from these regressions to verify that they were approximately normally distributed.

In constructing f, we let $f_1$ be the conditional distribution of ability and log income, given location of the student's home; $f_2$, the distribution of log distance. The $\mu_{ij}$'s were obtained from the regression equations. The $\Sigma_{ij}$'s were taken as the sample covariance matrices of residuals within state of residence, sex, and certain categories of institutions.[17]

STUDENT CHOICE PROBABILITIES

The problem is now to predict the institution an individual will choose, based on his vector of characteristics $Y_i$. Assume for the moment that there are K institutions on his list, which might include all institutions in the nation, or simply all institutions within a given distance radius. If we let P(j|i) be the probability that student i chooses institution j, Bayes' Theorem yields

$$P(j|i) = P(j)f(Y_i|Z_{ij}) \Big/ \sum_{k=1}^{K} P(k)f(Y_i|Z_{ik}) , \qquad (7)$$

where f is as above, and P(k) is the prior probability of choosing institution k.

[17] We observed that the dispersion of residuals for California two-year public and Massachusetts high-cost private institutions differed from the state-wide pattern; in these cases, we used their specific sample covariance matrices.

We took the prior probability P(k) to be proportional to the size of the freshman class by sex.[18] We felt that this was the best analysis-independent indicator of institutions' relative abilities to attract and absorb a student. It controls for an institution's capacity constraints. At the same time, it reflects the several factors (academic reputation, recruiting efforts, etc.) that affect student choices for which data are not available.

FEATURES OF THE APPROACH

Thus, a formula for the probability of an individual choosing a specific institution has been derived. According to Eqs. (5) and (6), the class of models is quite rich. And, unlike earlier models, this formula utilizes information not directly related to preference: institutional size, for example.

The approach succeeds in placing the task of modeling back into the familiar framework of ordinary linear regression, translating the problem of predicting choice into the problem of predicting characteristics. Thus, it is possible to utilize many of the important and familiar features of the linear model, including the ability to look at several different regressions based on one accumulation, the ability to test hypotheses about the effects of groups of variables, and the ability to examine lack of fit via residual plots. Computational costs are also orders of magnitude lower.

But the most important feature of the model is that it avoids the fundamental problem of imputing each student's choice set. Here, the alternative institutions only enter in defining independent variables. Thus, if the institution was not considered--and hence its particular attributes were unimportant--the corresponding independent variables will be expected to have coefficients close to zero. As an example,

---

[18] The Institutional Domain File provided these data for two-year colleges for the prior year, 1965. For other institutions, however, the data pertained to 1967, the year after the SCOPE students matriculated. Since the SCOPE students comprised only a small fraction of total enrollments in 1966, we assumed that 1967 enrollments were independent of SCOPE students' choices and, thus, that they could be used in Bayes' formula.

it might be reasonable to suppose that a high ability student with a public university nearby would be less likely to enter a two-year college than a similar student with no public university nearby. If true, the ability of a student at a junior college will depend on the presence or absence of a public university near his home; that hypothesis could be investigated by including in the ability prediction equation the appropriate indicator variable.

## IV. MODEL ESTIMATION

In this section, we provide details of our empirical analysis of the distribution of student characteristics from the SCOPE and Institutional Domain File data bases.

### CONSTRUCTING THE DATA BASE

Based on the studies cited earlier, we assumed that student ability, family income, and location of residence were the important student characteristics; that institutional type, cost, and location were the important institutional attributes. In all, we were able to obtain complete records for some 14,851 of the original cases. Below, we describe briefly how the variables were constructed.

### Student Ability

SCOPE used the standardized achievement test (AAT) to obtain measures of student verbal and math achievement. Most students in the SCOPE sample took the test; we excluded those who did not take both parts. Initially, we treated the verbal and math scores separately, but we found no useful information in their joint distribution. In the end, we used the sum of the two test scores as a single measure of ability.

### Student's Family Income

We used the RM procedure for imputing family income, truncating their estimates to the interval \$5,000 to \$25,000.[19] This specification was broad enough so that an income figure could be imputed for all records.

### Student Residency Location

We obtained high school latitude and longitude for all but one high school. We reasoned that this would be a satisfactory approximation of

[19] RM fit a linear regression model which we believe gave poor estimates at the extremes.

students' places of residence. The exception, of course, would be students whose families had moved; but we had no information about movers, and we felt that their number would not be large enough to have a major impact on our results.

Institutional Quality

The Institutional Domain File contains the average Scholastic Aptitude Test score (math plus verbal) for students at each institution. Following KMM and RM, we use this as the measure of institutional quality.

Institutional Cost

We estimate institutional cost as follows:

[TUITION] + [ROOM & BOARD] + [COMMUTATION COSTS] .

*Tuitions* at public sector institutions were obtained from college catalogues; tuitions for the private institutions were obtained from the Institutional Domain File.

*Room and Board* was assumed to be zero if the institution was within 30 miles of his home; equating 30 miles with 50 minutes driving time would make this consistent with the RM study. For institutions farther away, we used the room and board fee provided by the Institutional Domain File if available; otherwise, we used the national average room and board fee of $972 for public institutions, $1,140 for private institutions.

*Commutation Costs*, again taken from RM, were assumed to be $180 for institutions within 30 miles of a student's home; zero otherwise.

ESTIMATING THE DISTRIBUTION OF STUDENT CHARACTERISTICS

Spurred on by what we thought was a rather large data base, we initially posed models of student characteristics that were rich in parameters, conditioning on a large number of aspects of the student's institution and home environments. The richer models tended to yield inconsistencies, usually in the form of counter-intuitive signs on

regression coefficients in certain strata of the SCOPE population. Our response was generally to look at simpler models that yielded plausible results, and the final equations, obtained after systematically eliminating the spurious fits, are fairly parsimonious. Only these final results are reported.

We began by stratifying the SCOPE population into the eight groups: state of residence by sex. Within each group, we conditioned on the location of high school and choice of institution, and attempted to model the student's joint ability and income distribution; then, conditioning only on the choice of institution, we attempted to model the location of students' high schools.

We divided the institutions available to a given student into five types: (1) public two-year colleges, (2) public four-year colleges, (3) public universities, (4) low-cost (tuition $\leq$ \$1,000) private institutions, and (5) high-cost (tuition > \$1,000) private institutions. We reasoned that the regression coefficients on institutional attributes would be likely to depend on some categorization such as this, and in forming our models, we interacted them separately with the various independent variables.

Ability

Table 2 shows the results of the ability regressions. The equations have institutional type main effects and quality by institutional type interactions. We note that there is significant variation in the coefficients within each equation: tests for the importance of the main effects and for the institutional quality interactions showed these terms to be significant. And, where coefficients of institutional quality are significantly different from zero, they generally have the right (positive) sign, consistent with higher quality schools attracting higher ability students.

Of course, in the present circumstances, it is very important to investigate whether the distribution of the residuals is normal--this would be a necessary condition for the distribution of student characteristics to be multivariate normal. Thus, we obtained a random sample of 200 observations and plotted residuals separately against predicted

Table 2

RESULTS OF STUDENT ABILITY REGRESSIONS

| Subsample | Sample Size | $R^2$ | Estimated Standard Deviation | Regression Coefficients | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Type of Institution Dummy | | | | Type of Institution, Quality Interactions[a] | | | | |
| | | | | Constant Term | Public 2-yr | Public 4-yr | Public University | Private (Tuition ≤ $1000) | Public 2-yr | Public 4-yr | Public University | Private (Tuition ≤ $1000) | Private (Tuition > $1000) |
| California | | | | | | | | | | | | | |
| Males | 2133 | 0.30 | 14.24 | 8.71 | 55.33 | 10.55 | 35.29 | 11.75 | -1.10 | 4.97 | 2.91 | 4.74 | 5.85 |
| | | | | (1.0)[b] | (4.9) | (0.7) | (2.7) | (0.4) | (1.3) | (4.2) | (3.3) | (1.6) | (7.8) |
| Females | 1898 | 0.32 | 13.09 | -6.43 | 56.63 | 61.26 | 33.23 | 3.06 | -0.08 | -0.74 | 3.92 | 6.24 | 6.63 |
| | | | | (0.7) | (5.0) | (2.9) | (2.8) | (0.1) | (0.1) | (0.4) | (5.3) | (2.8) | (8.3) |
| Illinois | | | | | | | | | | | | | |
| Males | 2209 | 0.30 | 12.49 | 4.04 | -16.87 | -2.92 | 22.68 | 63.45 | 7.75 | 6.46 | 4.31 | -0.27 | 6.25 |
| | | | | (0.8) | (1.5) | (0.3) | (3.3) | (5.9) | (7.0) | (6.3) | (9.4) | (0.3) | (14.0) |
| Females | 1810 | 0.25 | 12.69 | -4.38 | 13.30 | 41.64 | 25.14 | 38.19 | 5.00 | 2.43 | 4.44 | 3.15 | 6.82 |
| | | | | (0.8) | (0.8) | (2.3) | (3.1) | (3.9) | (2.8) | (1.3) | (7.8) | (3.8) | (12.8) |
| Massachusetts | | | | | | | | | | | | | |
| Males | 1700 | 0.37 | 11.34 | 19.90 | 74.56 | -36.52 | 6.55 | -11.70 | -4.46 | 8.42 | 4.22 | 5.82 | 4.72 |
| | | | | (5.7) | (2.5) | (3.6) | (0.4) | (0.7) | (1.3) | (9.0) | (3.3) | (2.9) | (15.6) |
| Females | 1255 | 0.42 | 10.99 | 1.19 | 115.58 | 30.24 | 18.61 | 15.11 | -7.61 | 3.29 | 4.71 | 4.26 | 6.23 |
| | | | | (0.3) | (3.0) | (1.9) | (1.1) | (1.1) | (1.7) | (2.1) | (3.1) | (3.0) | (18.1) |
| North Carolina | | | | | | | | | | | | | |
| Males | 1989 | 0.50 | 11.30 | 13.86 | 33.58 | -20.78 | 20.08 | -12.43 | ---[c] | 7.64 | 3.72 | 6.46 | 5.31 |
| | | | | (1.6) | (3.9) | (2.2) | (1.3) | (1.3) | | (18.7) | (3.3) | (14.2) | (7.3) |
| Females | 1857 | 0.49 | 10.97 | 7.14 | 36.57 | -27.50 | -46.15 | 15.89 | ---[c] | 8.69 | 9.95 | 7.28 | 5.77 |
| | | | | (0.8) | (4.1) | (2.9) | (2.2) | (1.7) | | (27.0) | (5.6) | (18.7) | (7.0) |

[a] Institutional quality is defined as the sum of the average math and verbal Scholastic Aptitude Test scores, divided by 100.

[b] t statistics are shown in parentheses.

[c] No variation in the independent variable; variable omitted from the regressions.

values, institutional type, and quality to look for departures from the homoscedastic patterns; we also looked at normal probability plots of the residuals. We concluded in all cases that the residuals looked fairly normal, but in two instances (California public two-year colleges, Massachusetts private high-cost institutions) the spread of the residuals for both males and females was larger than for the rest of the state. In these cases, we chose to fit separate variance terms to the ability residuals.

Income

We observed by looking at probability plots of various income regressions that the normal assumptions would be seriously violated unless income were transformed. The logarithmic transformation seemed to work reasonably well; we ended up using it exclusively throughout.

Table 3 provides the results of regressing log (income) on institutional type and institutional type interacted with cost. The coefficients of cost generally had the correct sign: where significant, they suggested that higher income students attended the more expensive schools. We found, however, that knowing institutional cost did not reduce the variance of log income by a large amount.

The normal probability plots of the log (income) residuals showed this variable to be approximately normal. It also appeared that the spread of the residuals was independent of the various independent variables.

Joint Distribution of Ability and Income

A final step in characterizing the distribution of these quantities was to investigate their joint distribution. The basic requirements for the use of Eqs. (5) and (6) (Sec. III) is that the residuals of the previous regressions should appear to have a bivariate normal distribution. We looked at scatterplots of ability and log (income) residuals within state and sex to see if, in fact, they formed an elliptical pattern. We observed no obvious violations in these scatterplots and concluded that the multivariate normal assumption for ability and log (income) was reasonably consistent with the data.

Table 3

RESULTS OF STUDENT LOG (INCOME) REGRESSIONS

| Subsample | Sample Size | $R^2$ | Estimated Standard Deviation | Regression Coefficients | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Constant Term | Type of Institution Dummy | | | | Type of Institution, Cost Interactions[a] | | | | |
| | | | | | Public 2-yr | Public 4-yr | Public University | Private (Tuition ≤ $1000) | Public 2-yr | Public 4-yr | Public University | Private (Tuition ≤ $1000) | Private (Tuition > $1000) |
| California | | | | | | | | | | | | | |
| Males | 2133 | 0.07 | 0.1237 | 4.0209 (108.8)[b] | 0.0299 (0.8) | 0.0362 (0.9) | 0.0594 (1.5) | -0.0003 (0.0) | 0.0287 (1.2) | 0.0451 (2.4) | 0.0462 (3.0) | 0.0438 (0.7) | 0.0590 (3.5) |
| Females | 1898 | 0.10 | 0.1196 | 4.1300 (139.9) | -0.0788 (2.6) | -0.0738 (2.3) | -0.0365 (1.1) | -0.1537 (2.5) | 0.0350 (1.7) | 0.0631 (3.1) | 0.0456 (3.5) | 0.0861 (2.3) | 0.0118 (0.9) |
| Illinois | | | | | | | | | | | | | |
| Males | 2209 | 0.14 | 0.1179 | 3.9111 (167.8) | 0.1040 (4.2) | 0.1195 (4.0) | 0.0730 (2.8) | 0.1258 (3.2) | -0.0058 (0.3) | 0.276 (1.6) | 0.0771 (8.7) | 0.0125 (0.7) | 0.0932 (8.8) |
| Females | 1810 | 0.16 | 0.1126 | 3.9529 (200.4) | 0.0721 (3.3) | 0.0512 (2.0) | 0.0359 (1.6) | 0.0858 (1.9) | -0.0047 (0.2) | 0.0478 (2.8) | 0.0902 (9.3) | 0.0290 (1.1) | 0.0728 (8.5) |
| Massachusetts | | | | | | | | | | | | | |
| Males | 1700 | 0.10 | 0.1220 | 3.9176 (197.6) | 0.814 (3.2) | 0.0712 (2.6) | 0.0977 (3.6) | 0.0856 (1.9) | 0.0511 (1.7) | 0.0321 (1.7) | 0.0378 (2.7) | 0.0225 (0.8) | 0.0765 (8.7) |
| Females | 1255 | 0.18 | 0.1092 | 3.9436 (210.6) | 0.0852 (3.2) | 0.0307 (1.1) | 0.0240 (0.8) | 0.0533 (1.1) | -0.0009 (0.0) | 0.0550 (2.5) | 0.0832 (5.2) | 0.0335 (1.1) | 0.0702 (9.2) |
| North Carolina | | | | | | | | | | | | | |
| Males | 1989 | 0.09 | 0.1257 | 4.1450 (71.4) | -0.1389 (2.3) | -0.1495 (2.5) | -0.0985 (1.5) | -0.1951 (3.2) | 0.0073 (0.3) | 0.0297 (1.7) | -0.0338 (1.0) | 0.0684 (5.3) | 0.0014 (0.0) |
| Females | 1857 | 0.10 | 0.1187 | 3.9896 (82.5) | 0.0228 (0.5) | 0.0044 (0.1) | 0.1714 (2.5) | -0.0645 (1.3) | -0.0160 (0.5) | 0.0296 (1.5) | -0.0498 (1.4) | 0.0817 (7.3) | 0.0597 (3.0) |

[a]Cost is measured in units of $1000.

[b]t statistics are shown in parentheses.

Location of High School

We assume that the distribution of the location of high school was a function of the distance between the high school and institution. As with income, distance was transformed to logarithms; then, a simple model was fit including only dummy variables for institutional type. The equations are shown in Table 4. Our search for heteroscedastic or nonnormal patterns in the residuals proved negative, and we concluded that the normality assumptions were approximately true.

Table 4

RESULTS OF STUDENT LOG (DISTANCE) REGRESSIONS

| Subsample | Sample Size | $R^2$ | Estimated Standard Deviation | Regression Coefficients | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | Constant Term | Type of Institution Dummy | | | |
| | | | | | Public 2-yr | Public 4-yr | Public University | Private (Tuition ≤ $1000) |
| California | | | | | | | | |
| Males | 2133 | 0.42 | 0.6139 | 1.910 (38.6)[a] | -1.346 (25.9) | -0.497 (8.0) | -0.314 (5.0) | 0.402 (3.6) |
| Females | 1898 | 0.41 | 0.6763 | 1.955 (39.5) | -1.359 (25.4) | -0.517 (7.7) | -0.388 (6.2) | 0.291 (3.1) |
| Illinois | | | | | | | | |
| Males | 2209 | 0.38 | 0.6516 | 1.840 (62.6) | -1.173 (29.1) | -0.045 (0.9) | 0.40 (1.0) | -0.024 (0.5) |
| Females | 1810 | 0.40 | 0.6455 | 1.919 (63.5) | -1.366 (31.3) | -0.262 (5.6) | -0.196 (4.6) | -0.061 (1,0) |
| Massachusetts | | | | | | | | |
| Males | 1700 | 0.16 | 0.7351 | 1.504 (56.2) | -0.682 (13.5) | -0.385 (7.9) | 0.360 (6.5) | -0.121 (1.5) |
| Females | 1255 | 0.24 | 0.6858 | 1.703 (57.3) | -0.981 (17.0) | -0.609 (12.2) | 0.053 (0.8) | -0.191 (2.3) |
| North Carolina | | | | | | | | |
| Males | 1989 | 0.33 | 0.6603 | 2.095 (32.8) | -1.461 (19.8) | -0.508 (7.3) | -0.025 (0.4) | -0.586 (8.4) |
| Females | 1857 | 0.20 | 0.6710 | 1.993 (33.7) | -1.251 (16.6) | -0.312 (4.9) | 0.125 (1.3) | -0.473 (7.3) |

[a] t statistics are shown in parentheses.

## V. PREDICTIVE POWER

This section reviews RM's tests of the predictive accuracy of their model, reports the results of a similar test of ours, and compares the two sets of results.[20]

### RM'S SIMULATIONS

RM drew two samples of students: Sample I consists of 369 students whose parents had not responded to the family income question; Sample II consists of 375 students whose parents had reported family income. Each sample contains approximately equal numbers of students from each state. They further divided each sample by student test scores into four ability groups. Then they used estimated family income in all Sample I analyses, but performed Sample II analyses separately using parent reported income (IIA) and estimated income (IIB).

RM estimate the parameters of their model separately for each of the 12 cases (four ability groups by Samples I, IIA, and IIB). They calculate the probability that each student will choose each option in his choice set.[21] The probabilities are summed by option to obtain the predicted distribution of students among options.

To facilitate comparisons with our results, we eliminated predicted and actual nongoers, and rescaled the predicted and actual distributions of college goers to sum to one. RM's rescaled results are displayed in Tables 5 and 6.

### BAYESIAN SIMULATIONS

Our model can be used to predict the distribution of students over all colleges in the country. However, the predicted probability that a student will attend any particular college rapidly declines with distance.

---

[20] KMM did not provide a test of the predictive accuracy of their model.

[21] Recall, in RM's formulation, that a student's choice set consists of not going on to college or attending one of nine "representative" institutions, each of which offers the mean attributes of the institutions in a cost by quality category.

Table 5

RADNER AND MILLER STUDY RESULTS, SAMPLE I: PERCENTAGE DISTRIBUTION OF STUDENTS ATTENDING INSTITUTIONS IN COST/QUALITY CATEGORIES, BY STUDENTS' ABILITY GROUP[a]

| Institution Category | | Low Ability Group | | Medium Low Ability Group | | Medium High Ability Group | | High Ability Group | | All Ability Groups | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Quality[b] | Cost[c] | Predicted | Actual | Predicted | Actual | Predicted | Actual | Predicted | Actual | Predicted | Actual |
| Low | Low | 64.5 | 56.8 | 35.8 | 43.9 | 21.6 | 16.7 | 10.7 | 15.8 | 33.8 | 32.9 |
| Low | Medium | 23.0 | 16.2 | 38.6 | 21.9 | 12.1 | 14.6 | 7.6 | 2.6 | 19.7 | 12.4 |
| Low | High | 0.1 | 2.7 | 0.7 | 7.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 2.5 |
| Medium | Low | 3.8 | 5.4 | 3.7 | 4.9 | 18.6 | 4.1 | 15.6 | 5.3 | 10.7 | 5.0 |
| Medium | Medium | 6.0 | 16.2 | 10.5 | 12.2 | 18.5 | 18.8 | 11.7 | 21.1 | 12.0 | 17.4 |
| Medium | High | 2.6 | 2.7 | 10.6 | 7.3 | 10.7 | 8.3 | 10.9 | 2.6 | 8.6 | 5.6 |
| High | Low | 0.0 | 0.0 | 0.0 | 2.5 | 3.6 | 2.1 | 9.0 | 2.6 | 3.0 | 1.9 |
| High | Medium | 0.0 | 0.0 | 0.0 | 0.0 | 12.4 | 25.0 | 22.5 | 15.8 | 8.6 | 11.2 |
| High | High | 0.0 | 0.0 | 0.2 | 0.0 | 2.5 | 10.4 | 12.0 | 34.2 | 3.4 | 11.2 |

SOURCE: Radner and Miller [2], Tables 25-28, pp. 106-107. Recompiled over nine institutional categories.

[a]Low = less than 400; medium low = 400-475; medium high = 475-550; high = 550+.

[b]Measure of quality = average SAT score of all students attending the institution. Low = less than 480; medium = 480-540; high = 540+.

[c]Low = less than $600 per year; medium = $600-$2250 per year; high = $2250+ per year.

Table 6

RADNER AND MILLER STUDY RESULTS, SAMPLE II: PERCENTAGE DISTRIBUTION OF STUDENTS ATTENDING INSTITUTIONS IN COST/QUALITY CATEGORIES FOR ALTERNATIVE MEASURES OF STUDENTS' FAMILY INCOME, BY STUDENTS' ABILITY GROUP[a]

| Institution Category | | Low Ability Group | | | Medium Low Ability Group | | | Medium High Ability Group | | | High Ability Group | | | All Ability Groups | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Quality[b] | Cost[c] | Parent-reported Income | Estimated Income | Actual | Parent-reported Income | Estimated Income | Actual | Parent-reported Income | Estimated Income | Actual | Parent-reported Income | Estimated Income | Actual | Parent-reported Income | Estimated Income | Actual |
| Low | Low | 47.4 | 40.5 | 39.4 | 24.9 | 30.7 | 43.3 | 22.3 | 25.2 | 30.5 | 5.8 | 6.4 | 9.6 | 27.2 | 27.4 | 32.1 |
| Low | Medium | 31.8 | 38.0 | 32.4 | 20.4 | 15.2 | 22.4 | 22.1 | 19.7 | 22.0 | 6.0 | 5.9 | 3.9 | 21.2 | 21.5 | 21.3 |
| Low | High | 0.6 | 1.3 | 5.6 | 0.4 | 0.1 | 1.5 | 0.3 | 0.2 | 3.4 | 0.1 | 0.0 | 1.9 | 0.4 | 0.5 | 3.2 |
| Medium | Low | 3.3 | 2.0 | 8.4 | 11.8 | 21.1 | 7.5 | 9.2 | 12.5 | 6.8 | 7.4 | 8.0 | 9.6 | 7.6 | 10.2 | 8.0 |
| Medium | Medium | 7.9 | 7.4 | 7.1 | 19.8 | 18.9 | 16.4 | 16.9 | 16.1 | 20.3 | 13.8 | 14.5 | 17.3 | 14.1 | 13.6 | 14.9 |
| Medium | High | 8.9 | 10.6 | 5.6 | 20.4 | 11.0 | 3.0 | 17.3 | 11.9 | 1.7 | 18.0 | 15.0 | 5.8 | 15.6 | 11.9 | 4.0 |
| High | Low | 0.0 | 0.0 | 0.0 | 0.4 | 0.8 | 1.5 | 2.6 | 4.1 | 6.8 | 4.5 | 5.9 | 0.0 | 1.6 | 2.4 | 2.0 |
| High | Medium | 0.0 | 0.0 | 0.0 | 1.1 | 1.7 | 1.5 | 6.8 | 8.5 | 5.1 | 21.0 | 23.4 | 15.4 | 6.4 | 7.4 | 4.8 |
| High | High | 0.1 | 0.1 | 1.4 | 0.9 | 0.5 | 3.0 | 2.5 | 1.9 | 2.3 | 23.5 | 20.9 | 36.5 | 6.0 | 5.1 | 9.6 |

SOURCE: Radner and Miller [2], Tables 25-28, pp. 106-107. Recompiled over nine institutional categories.

[a]Low = less than 400; medium low = 400-475; medium high = 475-550; high = 550+.

[b]Measure of quality = average SAT score of all students attending the institution. Low = less than 480; medium = 480-540; high = 540+.

[c]Low = less than $600 per year; medium = $600-$2250 per year; high = $2250+ per year.

We felt that a simulation of students' choices among the institutions "near" his home would lead to reasonably accurate predictions and would be much less expensive; arbitrarily, we chose a 50 mile boundary.

We use Eqs. (5) through (7) to calculate the probability that each student would attend each institution located within 50 miles of his high school; the 155 students with no institution within 50 miles were deleted. We then stratified the institutions into RM's nine quality/cost categories, and summed the estimated probabilities over institutions in each category. Finally, we counted the actual number of students in each category, regardless of whether they attended an institution within 50 miles.

Table 7 shows the predicted and actual number of students in each state who attended institutions in each of the nine categories. We then stratified the students by the RM ability criteria and summed over states to obtain the predicted and actual number of students in each ability group by institutional category. Table 8 presents these data in the format of Tables 5 and 6, facilitating a comparison of our results with those of RM.

## COMPARISON OF RESULTS

We used the Gini coefficient [10] to measure the accuracy of the predicted frequency distributions. It is the sum of the absolute differences between the predicted and actual frequencies; higher values thus imply greater discrepancies between these distributions. Table 9 provides Gini coefficients for each of the simulations discussed above. It is clear that our predicted distributions are substantially closer than RM's to the actual distributions in every case.

We recognized, however, that according to the law of large numbers, this comparison favored the Bayes approach: it utilized more than 14,000 observations whereas RM used fewer than 400. So, we randomly assigned each of the 14,696 students in our sample to one of 40 subsamples, and replicated the simulation in each case.[22] We computed the Gini coefficients for the predicted and corresponding actual distributions of

[22] Subsample sizes ranged from 335 to 405, averaging 367.

Table 7

PREDICTED AND ACTUAL NUMBER OF STUDENTS ATTENDING INSTITUTIONS, BY COST/QUALITY CATEGORIES AND BY STATE

| Institution Category[a] | State | Cost Category: Low (Less than $600 per year) | | Medium ($600-$2250 per year) | | High ($2250+ per year) | | All | |
|---|---|---|---|---|---|---|---|---|---|
| | | Predicted | Actual | Predicted | Actual | Predicted | Actual | Predicted | Actual |
| Low (Less than 480) | Calif. | 2174 | 2425 | 320 | 154 | 23 | 12 | 2517 | 2591 |
| | Ill. | 1581 | 1309 | 582 | 732 | 17 | 46 | 2180 | 2087 |
| | Mass. | 238 | 457 | 562 | 321 | 90 | 78 | 889 | 856 |
| | N.C. | 905 | 642 | 1965 | 1736 | 57 | 29 | 2927 | 2407 |
| | Total | 4898 | 4833 | 3429 | 2943 | 187 | 165 | 8514 | 7941 |
| Medium (480-540) | Calif. | 535 | 318 | 380 | 477 | 25 | 30 | 940 | 825 |
| | Ill. | 170 | 93 | 957 | 946 | 39 | 96 | 1166 | 1135 |
| | Mass. | 37 | 52 | 549 | 730 | 55 | 112 | 642 | 894 |
| | N.C. | 54 | 192 | 397 | 436 | 16 | 20 | 467 | 648 |
| | Total | 796 | 655 | 2283 | 2589 | 135 | 258 | 3214 | 3502 |
| High (540+) | Calif. | 243 | 247 | 241 | 265 | 76 | 90 | 561 | 602 |
| | Ill. | 13 | 12 | 524 | 520 | 136 | 265 | 673 | 797 |
| | Mass. | 13 | 13 | 998 | 766 | 413 | 426 | 1424 | 1205 |
| | N.C. | 0 | 0 | 294 | 539 | 17 | 110 | 311 | 649 |
| | Total | 269 | 272 | 2057 | 2090 | 642 | 891 | 2968 | 3253 |
| All | Calif. | 2952 | 2990 | 941 | 896 | 125 | 132 | 4018 | 4018 |
| | Ill. | 1764 | 1414 | 2063 | 2198 | 191 | 407 | 4019 | 4019 |
| | Mass. | 288 | 522 | 2109 | 1817 | 558 | 616 | 2955 | 2955 |
| | N.C. | 959 | 834 | 2656 | 2711 | 90 | 159 | 3705 | 3704 |
| | Total | 5963 | 5760 | 7769 | 7622 | 964 | 1312 | 14696 | 14696 |

[a]Measure of quality - average SAT score of all students attending the institution.

Table 8

RAND STUDY RESULTS: PERCENTAGE DISTRIBUTION OF STUDENTS ATTENDING INSTITUTIONS IN COST/QUALITY CATEGORIES, BY STUDENTS' ABILITY GROUP[a]

| Institution Category | | Low Ability Group | | Medium Low Ability Group | | Medium High Ability Group | | High Ability Group | | All Ability Groups | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Quality[b] | Cost[c] | Predicted | Actual | Predicted | Actual | Predicted | Actual | Predicted | Actual | Predicted | Actual |
| Low | Low | 52.1 | 56.3 | 37.2 | 36.0 | 25.3 | 23.0 | 14.1 | 10.6 | 33.3 | 32.9 |
| Low | Medium | 30.3 | 27.1 | 28.3 | 25.6 | 21.5 | 17.8 | 11.2 | 7.4 | 23.3 | 20.0 |
| Low | High | 1.5 | 1.5 | 1.8 | 1.3 | 1.2 | 1.2 | 0.4 | 0.4 | 1.3 | 1.1 |
| Medium | Low | 4.7 | 2.7 | 5.3 | 5.0 | 6.0 | 6.1 | 5.7 | 4.3 | 5.4 | 4.5 |
| Medium | Medium | 8.1 | 9.0 | 16.1 | 20.0 | 20.5 | 24.5 | 18.7 | 18.2 | 15.5 | 17.6 |
| Medium | High | 0.5 | 0.7 | 1.1 | 2.1 | 1.2 | 2.5 | 1.0 | 1.9 | 0.9 | 1.8 |
| High | Low | 0.6 | 0.3 | 1.0 | 1.0 | 2.1 | 2.2 | 4.2 | 4.3 | 1.8 | 1.8 |
| High | Medium | 1.7 | 1.9 | 7.1 | 7.1 | 17.2 | 16.9 | 33.7 | 34.8 | 14.0 | 14.2 |
| High | High | 0.4 | 0.4 | 2.0 | 2.0 | 5.1 | 5.8 | 11.1 | 18.1 | 4.4 | 6.1 |

[a]Low = less than 400; medium low = 400-475; medium high = 475-550; high = 550+.

[b]Measure of quality = average SAT score of all students attending the institution. Low = less than 480; medium = 480-540; high = 540+.

[c]Low = less than $600 per year; medium = $600-$2250 per year; high = $2250+ per year.

Table 9

COMPARISON OF STUDY RESULTS: ABSOLUTE VALUES OF DEVIATIONS BETWEEN PREDICTED AND ACTUAL PERCENTAGE DISTRIBUTIONS, BY STUDENTS' ABILITY GROUP

| Students' Ability Group[a] | Radner and Miller Study | | | Rand Study |
|---|---|---|---|---|
| | Sample I | Sample II | | |
| | | Parent Reported Income | Estimated Income | |
| Low | 29.0 | 24.1 | 24.0 | 11.0 |
| Medium low | 40.3 | 50.2 | 118.7 | 9.6 |
| Medium high | 46.6 | 38.9 | 37.4 | 12.5 |
| High | 73.4 | 48.7 | 50.2 | 18.3 |
| All | 36.1 | 26.2 | 26.5 | 9.7 |

[a]Low = less than 400; medium low = 400-475; medium high = 475-550; high = 550+.

students at each ability level and across ability levels. Table 10 shows the maximum, mean, minimum, and standard deviation of these Gini coefficients by student ability level. For reference purposes, it also shows smallest Gini coefficients for the three comparable RM predictions.

Our least accurate prediction, over 40 samples, is superior[23] to RM's most accurate prediction, over three samples, for students in the medium-low, medium-high, and high ability groups and across ability groups. In the case of low ability students, 34 of our 40 predicted distributions were more accurate than RM's most accurate prediction.

[23]That is, it had a lower Gini coefficient.

Table 10

DISTRIBUTION OF ABSOLUTE VALUES OF DEVIATIONS BETWEEN PREDICTED AND ACTUAL PERCENTAGE DISTRIBUTIONS FOR 40 INDEPENDENT SUBSAMPLES, BY STUDENTS' ABILITY GROUP

| Students' Ability Group[a] | Summary Statistics for 40 Subsamples | | | | |
|---|---|---|---|---|---|
| | Maximum | Mean | Minimum | Standard Deviation | Lowest Coefficient for 3 RM Samples |
| Low | 32.3 | 17.2 | 7.8 | 6.5 | 24.0 |
| Medium low | 35.6 | 20.7 | 9.7 | 6.4 | 40.3 |
| Medium high | 37.2 | 23.5 | 10.8 | 6.4 | 37.4 |
| High | 47.8 | 28.9 | 10.7 | 7.6 | 48.7 |
| All | 19.4 | 13.6 | 7.6 | 3.4 | 26.2 |

[a]Low = less than 400; medium low = 400-475; medium high = 475-550; high = 550+.

## VI. CONCLUDING REMARKS

We find that our predictions are considerably closer to the actual values than those based on the conditional logit approach. In addition, the Bayesian methodology is easier to use, offers much greater flexibility, and is much less expensive to apply. Thus, we feel that it offers considerable advantage over the conditional logit approach in the present context.

A number of recent studies have employed the conditional logit approach to model choice behavior in various areas, including education, transportation [6], and occupation [7,8]. While the Bayesian formulation might not be superior in all instances, our results suggest that it is a viable alternative. Even in those cases, where the conditional logit approach might be preferred on theoretical grounds, the Bayesian methodology would be a useful adjunct in the exploratory stages of research.

## REFERENCES


1. Kohn, Meir G., Charles F. Manski, and David S. Mundel, *An Empirical Investigation of Factors Which Influence College-Going Behavior*, The Rand Corporation, R-1470-NSF, September 1974.

2. Radner, Roy, and Leonard S. Miller, *Demand and Supply in U.S. Higher Education*, McGraw-Hill, New York, 1975.

3. Miller, Leonard S., and Roy Radner, *Demand and Supply in U.S. Higher Education: A Technical Supplement*, McGraw-Hill, New York, 1975.

4. McFadden, Daniel, "The Revealed Preferences of a Government Bureaucracy," Economic Growth Project, University of California, Berkeley, Technical Report No. 17, 1968.

5. -----, "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka (ed.), *Frontiers in Econometrics*, Academic Press, New York, 1973.

6. Charles River Associates Incorporated, *A Disaggregated Behavioral Model of Urban Travel Demand*, Final Report to U.S. Department of Transportation, Contract No. FH-11-7566, March 1972.

7. Schmidt, Peter, and Robert P. Strauss, "The Prediction of Occupation Using Multiple Logit Models," *International Economic Review*, Vol. 16, June 1975, pp. 471-486.

8. Boskin, Michael J., "A Conditional Logit Model of Occupational Choice," *Journal of Political Economy*, Vol. 82, March/April 1974, pp. 389-398.

9. Creager, John A., and Charles L. Sell, *The Institutional Domain of Higher Education: A Characteristics File for Research*, American Council on Education, Research Report, 1969.

10. Goodman, Leo A., and William H. Kruskal, "Measures of Association for Cross-Classifications. II: Further Discussion and References," *Journal of the American Statistical Association*, Vol. 54, March 1959, pp. 123-163.